Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W313–W317***
DOI: 10.1093/nar/gkh406

GOblet: a platform for Gene Ontology annotation of anonymous sequence data

Detlef Groth, Hans Lehrach and Steffen Hennig*

Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Received February 24, 2004; Revised and Accepted March 29, 2004

ABSTRACT

GOblet is a comprehensive web server application providing the annotation of anonymous sequence data with Gene Ontology (GO) terms. It uses a variety of different protein databases (human, murines, invertebrates, plants, sp-trembl) and their respective GO mappings. The user selects the appropriate database and alignment threshold and thereafter submits single or multiple nucleotide or protein sequences. Results are shown in different ways, e.g. as survey statistics for the main GO categories for all sequences or as detailed results for each single sequence that has been submitted. In its newest version, GOblet allows the batch submission of sequences and provides an improved display of results with the aid of Java applets. All output data, together with the Java applet, are packed to a downloadable archive for local installation and analysis. GOblet can be accessed freely at http://goblet.molgen.mpg.de.

INTRODUCTION

In recent years the Gene Ontology (GO) Consortium (1) has developed a rich and comprehensive unified vocabulary for the description of genes, their functions and their products, divided into the three main categories: 'Molecular function', 'biological process' and 'cellular component'. Currently the vocabulary comprises around 16 000 GO terms and is frequently updated, which reflects the ongoing activities and further improvements of the GO system (2). Nevertheless, it is widely used already in large-scale gene/protein annotation projects (3–8) and can be regarded as a standard for computer-based sequence annotation systems. The advantage of having a well-defined and standardized vocabulary for the description of genes is obvious: specific gene sets (e.g. tissues) or even complete genomes can easily be clustered and compared with respect to common functional features, and genome databases such as MGD (9) and GoFish (10) can be screened via complex Boolean queries based on GO terms (e.g. show me all genes involved in 'immunity' and 'cell–cell signaling'). A very interesting application of GO is in the analysis of gene expression experiments, where it has been used e.g. for comparison of different probe/chip sets (11) and for identification of transcriptional signatures conserved between distant species (12). Meanwhile, commercial oligo chips are delivered with corresponding GO annotations for the probes (13).

The species-independent GO vocabulary and hierarchy of terms are actively used by several groups to provide the scientific community with gene and protein sets annotated using GO terms (4,5,8,9). Currently the GO-based annotations (GOAs) available from the GO website (http://www.geneontology.org) comprise approximately 8 million associations and 1.6 million different gene products (2), with the largest set covering around 780 000 UniProt proteins (4) from a variety of species. However, for scientists with little computational background or without appropriate facilities it is a tedious task to annotate genes or proteins not yet included in the public GOA sets with GO terms. Our GOblet server was designed to aid GO annotation of anonymous sequences (cDNA, protein) based on similarity searches against a collection of specially designed protein databases (DBs). Since its first publication in 2003 (14) the service has been frequently used, and we now present a highly improved version which allows batch processing and has a newly designed interface for efficient browsing of analysis results and GOblet database entries.

THE STRUCTURE OF THE GOblet SYSTEM—DATABASES AND PRINCIPLES

The details of database construction and sequence analysis behind GOblet have been described elsewhere (14) and will be sketched only briefly.

(i) Public GO annotations are downloaded from the GO website or the Ensembl website (http://www.ensembl.org).
(ii) Corresponding sequence sets are transferred from various sources (see http://goblet.molgen.mpg.de for details) and compiled into a collection of databases which can be queried by BLAST (15,16).

*To whom correspondence should be addressed. Tel: +49 30 8413 1612; Fax: +49 30 8413 1380; Email: hennig@molgen.mpg.de

Table 1. List of protein databases available for GOblet analysis

Database	Description	Number of proteins	Number of GO-ids	Non-IEA
swiss-tr	GO annotations for proteins (all species) from SwissProt and TrEMBL (EBI)	810 379	7297	4999
swiss-tr-human	GO annotations for human proteins from SwissProt, TrEMBL (EBI)	27 924	4206	3106
ensembl-human	GO-ids for human proteins (ENSEMBL)	16 290	4970	4334
murines	The murine subset of SP-TrEMBL, mainly *Mus musculus* and *Rattus norvegicus* (EBI)	32 814	3559	2284
drosophila	GO annotations for *Drosophila melanogaster* proteins (flybase)	10 977	3489	3376
wormpep	GO annotated *Caenorhabditis elegans* transcripts (WormBase)	7406	917	25
yeast	GO annotated yeast transcripts (SGD)	5882	2372	2372
viridiplantae	The Viridiplantae (green plants) set is a subset of SP-TrEMBL	86 595	1676	78
a.thaliana	*Arabidopsis thaliana* GOA and protein sequences (TAIR)	29 013	2659	2098
o.sativa	Rice (*Oryzes sativa*) protein set (GRAMENE)	18 879	955	420

In all cases except human (ENSEMBL), the GO annotations were downloaded from the Gene Ontology website. Sequences corresponding to the GOA tables were extracted either from the SwissProt/TrEMBL database maintained at the European Bioinformatics Institute, or from species-specific repositories (TAIR, GRAMENE, SGD, WormBase, flybase). The last column gives the total amount of non-IEA evidence codes. Respective links to the source databases can be found at http://goblet.molgen.mpg.de.

(iii) Queries to GOblet (sequences) are uploaded via a web form, with the appropriate BLAST parameters (*E*-value, database) given by the user.
(iv) Automated parsers extract the relevant information from BLAST output (GO ids, target description, accession numbers, significance values).
(v) Visualization of results is via HTML documents which contain links to alignments, source documents (e.g. at SwissProt), GO terms and a summary tree of all GO terms associated with the query sequence.

The set of protein databases has been continuously expanded over the past year. As a result of requests from some users we have compiled plant- and murine-specific DB sets. Both sets were derived either from the SwissProt/TrEMBL protein set via the respective taxonomy information or from species-related public sources. A complete list of available DBs is given in Table 1.

GOblet was first published in July 2003. Since then, the service has been used by many users, who have performed, on average, nearly 1300 analysis runs per month. Until now searches have been limited to a single query sequence, but users of GOblet have been free to send many jobs sequentially. In any case each analysis result can later be traced through a unique URL displayed directly after submission. In the new version the GOblet server accepts batches of sequences. Query sequences have to be sent in one single file in Fasta format; data transfer is handled via cutting and pasting into an HTML form.

AN ADVANCED INTERFACE AND ANALYSIS SYSTEM FOR REMOTE AND LOCAL USE

Web applications are ideal solutions to overcome the typical difficulties of installing and maintaining various software packages on different platforms. However, the capabilities of the graphical user interfaces (GUIs) provided by standard web browsers are very limited. Therefore, we extended the previous GOblet interface by implementing a sophisticated Java applet. We chose Java (http://www.sun.java.com) because of its cross-platform availability and high flexibility in design and management of GUIs.

The new way that GOblet displays analysis results is based on the Thinlet toolkit (http://www.thinlet.com), which requires only a Java-enabled web browser. The Java applet can be run either as an applet inside the web browser or as a standalone application if an appropriate Java plugin or the Java Runtime Environment (JRE) is installed. The GUI is built using the XUL markup language (eXtensible User interface Language), which is used, for instance, within the famous Mozilla browser. Thinlet is a Java approach to rendering the GUI via this markup language. It has the technical advantages that the logic of the user interface is separated from the programming logic and that the GUI can be built either statically (via an XUL file included in the applet) or dynamically, e.g. by communication with a web server. This way it is possible to use the powerful Perl scripting language on the server side to deliver XUL data to the Java client (the user) for dynamic updates of the interface. Thinlet requires only a Java 1.1-enabled web browser, which minimizes problems due to old browser versions on the client side.

The main features of the upgraded GOblet system are summarized in the following; a detailed description is available online.

(i) Up to 100 KB of sequence data (nucleotide or protein) can be sent for a single analysis run, which corresponds to about 200 sequences with 500 bp length. The BLAST parameters have to be set by the user as described above. A major improvement over the last version of GOblet is the explicit use of evidence codes as provided by the GO consortium and associated databases. Evidence codes reflect the quality of annotation, with codes IEA (inferred by electronic annotation) and TAS (traceable author statement) being regarded as the lowest and highest level of confidence, respectively. GOblet users now have the option to exclude all IEA codes from the analysis in order to gain improved annotation quality. However, it should be noted that many public annotations have only a minor percentage of non-IEA codes, so that exclusion of IEA codes will reduce the search space considerably (Table 1). Immediately after submission, our server returns a valid URL which can be 'bookmarked' for later retrieval of the results.
(ii) The results are displayed in a compact window (Figure 1), with the lower part of the left frame showing the list of queries. Each query symbol is linked to result pages, which are designed in the same way as in the previous

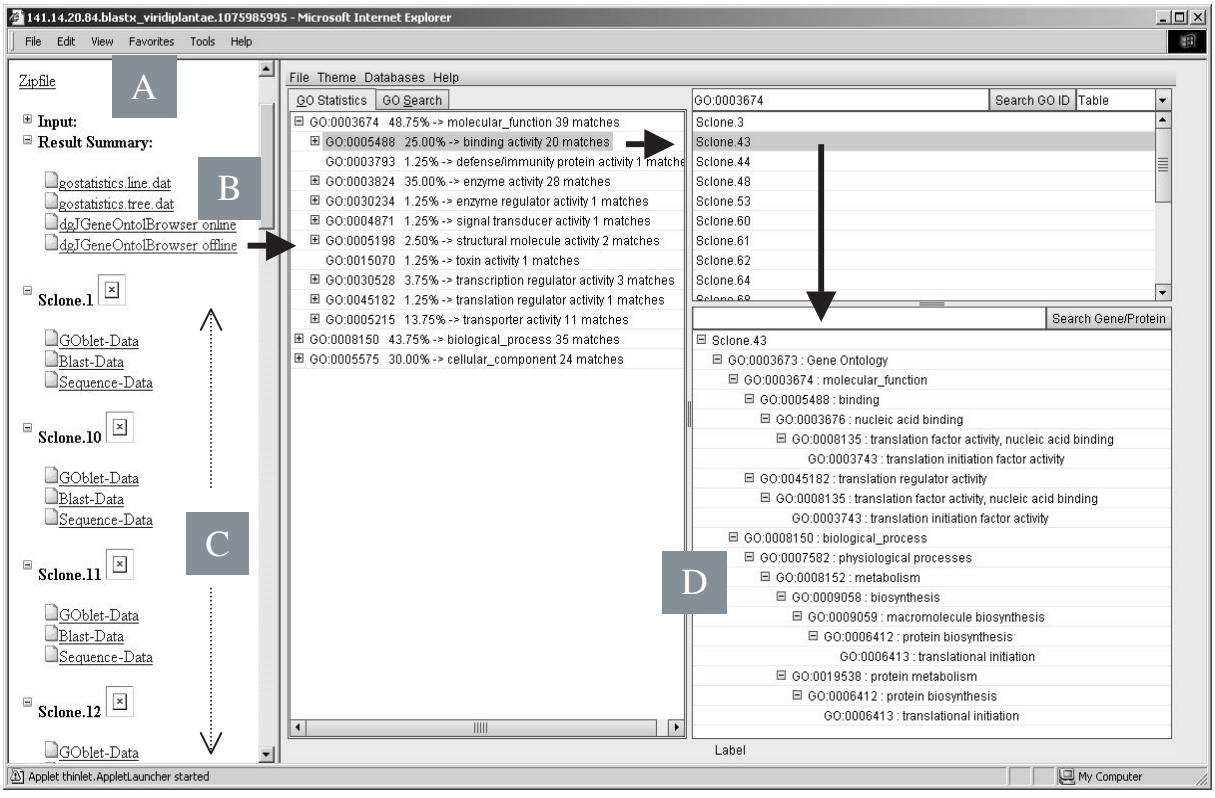

Figure 1. The main result window. (**A**) Zip archive for download and local installation of the complete set of analysis files and the java executable. (**B**) Summary results via links to the offline (local) or online (GOblet server) data sources. The two 'gostatistics' links provide displays of all query–GO id relationships and summary statistics in simple ASCII format for easy download. (**C**) Listing of all single query results. The links point to the query sequence, original BLAST output files and GOblet analysis page for each single query. (**D**) Frame controlled by the Java applet: a dynamic tree with summary statistics for the respective GO terms (nodes) is shown; clicking on a specific node opens the list of all query ids positive with that GO term. Selection of a specific query id in the list opens a summary tree for that query sequence.

version of GOblet (see above). In the upper part, links provide access to various analysis modes. Activation of the 'dgJGeneOntolBrowser' link starts the applet in the right-hand frame; a dynamical tree of GO terms is shown together with the percentage of respective genes/proteins falling into that class. 'Double clicking' on a GO class in the tree will produce a list of the respective members (genes/proteins). A detailed description is given in Figure 1. Results are stored on our server for at least one week.

(iii) An important and very convenient feature is the download function (Figure 1), which enables the user to store all results on a local system with a single click. Besides data, the archive contains all relevant tools to view the results offline. Once the archive is unpacked, clicking on the index.html file will open the same result window on the user's machine as in the online mode.

(iv) In the online mode it is possible to generate GO statistics for the databases hosted by our server. On selection of a specific database the dynamic GO tree is displayed via the Java applet, which provides the same features for the database entries as described above for the results of a GOblet analysis run (Figure 1). In addition, choosing the table mode allows comparison of different databases, e.g. of all plant protein sets available, as shown in Figure 2. The specificity of GO terms used for the statistics can be defined by the user. Since we found the comparative statistics a very useful tool, the GOblet server now provides an upload mechanism whereby users can send in their own GO associations. The server will store the respective IP address so that only connections via the same IP will be permitted to access the data thereafter.

DISCUSSION AND FUTURE DEVELOPMENTS

Annotation of genes and proteins using GO terms imported from homologous protein sequences is a very efficient approach to classifying not-yet-annotated sequences (14). Once the association with GO terms is built, novel sequence data can be easily classified with respect to functional groups, cellular compartments and general biological processes, with interesting applications in e.g. gene expression analysis (11,17,18). However, some care should be taken in the selection of the respective GOA set. The availability of several species- or phylum-specific databases allows the users of GOblet to restrict the search space to the appropriate protein sets; e.g. for most plant sequences the viridiplantae DB will be suitable, while for mammalian data the human or mouse DBs will be the best choice. The new version of GOblet provides processing of batches of sequences as well as user-friendly downloading of analysis results. In the future we will extend the service by including an email-based notification system; prior to job submission the user will be able to supply a valid email address at which to receive information about job completion and where to access the results. In addition,

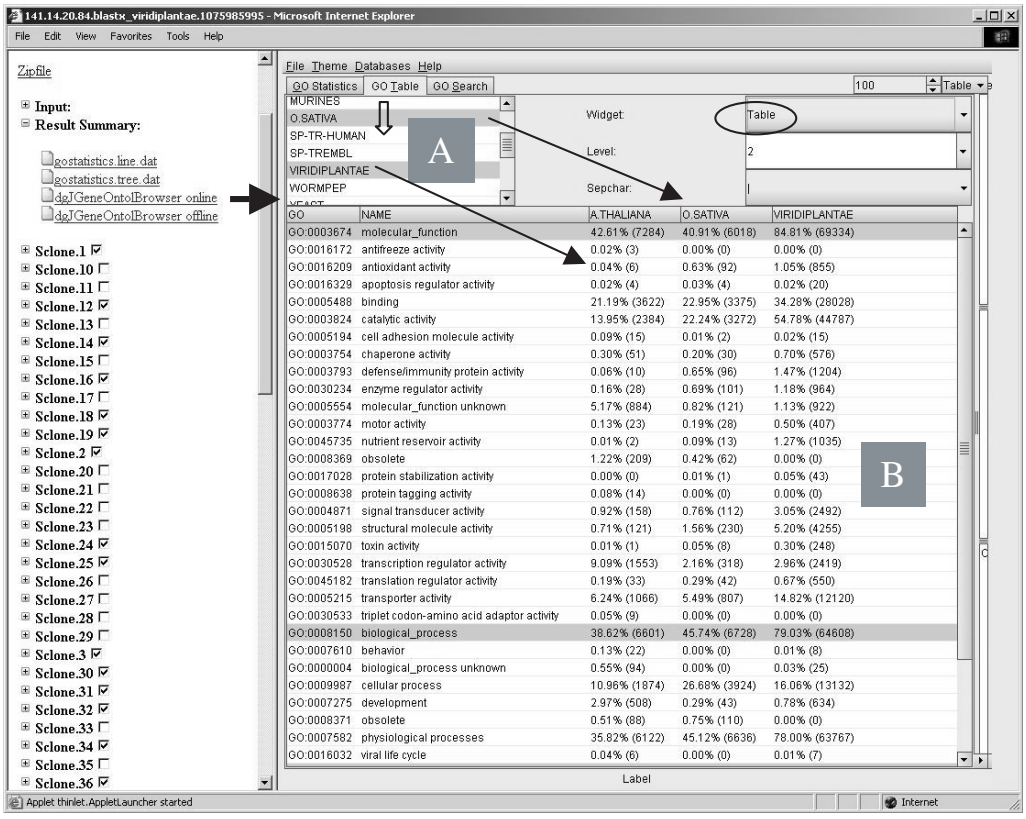

Figure 2. The comparative data view of the interface. (**A**) On selection of the 'GO Table' button a small menu pops up wherein several datasets can be selected. (**B**) After selection, activating the 'Table' field (top right) will give the percentages (numbers) of all genes/proteins falling into the respective GO classes. The specificity of classes can be controlled by the 'level' value (top right): level 1 refers to the three main classes of GO, level 2 contains the child terms of level 1, and so on. Note that GO associations for any set of genes or proteins can be uploaded from the user's end for comparative statistics.

the search engine of the server will be modified so that Boolean queries can be performed.

Due to the limitations of our server capacity there are still some restrictions on e.g. for the size of the query set and of the uploaded file of GO associations. However, scientists with very large datasets to be analysed are welcome to contact us about extended analysis.

AVAILABILITY AND TECHNICAL NOTES

GOblet is freely accessible at http://goblet.molgen.mpg.de. No special environments on the client side are required; any modern web browser such as Mozilla/Netscape, Opera or Internet Explorer can be used. However, to take full advantage of the Java applet the JRE needs to be installed, which should already have been done on most current platforms. On the server side is a DEC alpha workstation running the OSF1 operating system with an Apache web server. Perl/CGI scripts are executed to perform the communication between the user and the server and to process the results.

ACKNOWLEDGEMENTS

D.G. and S.H. thank the BMBF (German Ministry for Education and Research) for funding in the framework of the German National Genome Research Network (NGFN).

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
2. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
3. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662–672.
4. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
5. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
6. Poustka,A.J., Groth,D., Hennig,S., Thamm,S., Cameron,A., Beck,A., Reinhardt,R., Herwig,R., Panopoulou,G. and Lehrach,H. (2003) Generation, annotation, evolutionary analysis, and database integration of 20 000 unique sea urchin EST clusters. *Genome Res.*, **13**, 2736–2746.
7. Sprague,J., Clements,D., Conlin,T., Edwards,P., Frazer,K., Schaper,K., Segerdell,E., Song,P., Sprunger,B. and Westerfield,M. (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res.*, **31**, 241–243.
8. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism

database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.
9. Bult,C.J., Blake,J.A., Richardson,J.E., Kadin,J.A., Eppig,J.T., Baldarelli,R.M., Barsanti,K., Baya,M., Beal,J.S., Boddy,W.J. *et al*. (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nucleic Acids Res*., **32**, D476–D481.
10. Berriz,G.F., White,J.V., King,O.D. and Roth,F.P. (2003) GoFish finds genes with combinations of Gene Ontology attributes. *Bioinformatics*, **19**, 788–789.
11. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*., **31**, 3775–3781.
12. McCarroll,S.A., Murphy,C.T., Zou,S., Pletcher,S.D., Chin,C.S., Jan,Y.N., Kenyon,C., Bargmann,C.I. and Li,H. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genet*., **36**, 197–204.
13. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*., **31**, 82–86.
14. Hennig,S., Groth,D. and Lehrach,H. (2003) Automated Gene Ontology annotation for anonymous sequence data. *Nucleic Acids Res*., **31**, 3712–3715.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
17. Doniger,S.W., Salomonis,N., Dahlquist,K.D., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*., **4**, R7.
18. Zhong,S., Li,C. and Wong,W.H. (2003) ChipInfo: software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res*., **31**, 3483–3486.